Choose a Pod

Selecting the appropriate Pod instance is a critical step in planning your RunPod deployment. The choice of VRAM, RAM, vCPU, and storage, both Temporary and Persistent, can significantly impact the performance and efficiency of your project.

This page gives guidance on how to choose your Pod configuration. However, these are general guidelines. Keep your specific requirements in mind and plan accordingly.

Overview

It's essential to understand the specific needs of your model. You can normally find detailed information in the model card's description on platforms like Hugging Face or in the `config.json` file of your model.

There are tools that can help you assess and calculate your model's specific requirements, such as:

- Hugging Face's Model Memory Usage Calculator
- Vokturz' Can it run LLM calculator
- Alexander Smirnov's VRAM Estimator

Using these resources should give you a clearer picture of what to look for in a Pod.

When transitioning to the selection of your Pod, you should focus on the following main factors:

- **GPU**
- **VRAM**
- **Disk Size**

Each of these components plays a crucial role in the performance and efficiency of your deployment. By carefully considering these elements along with the specific requirements of your project as shown in your initial research, you will be well-equipped to determine the most suitable Pod instance for your needs.

GPU

The type and power of the GPU directly affect your project's processing capabilities, especially for tasks involving graphics processing and machine learning.

Importance

The GPU in your Pod plays a vital role in processing complex algorithms, particularly in areas like data science, video processing, and machine learning. A more powerful GPU can significantly speed up computations and enable more complex tasks.

Selection Criteria

- **Task Requirements**: Assess the intensity and nature of the GPU tasks in your project.
- **Compatibility**: Ensure the GPU is compatible with your software and frameworks.
- **Energy Efficiency**: Consider the power consumption of the GPU, especially for long-term deployments.

VRAM

VRAM (Video RAM) is crucial for tasks that require heavy graphical processing and rendering. It is the dedicated memory used by your GPU to store image data that is displayed on your screen.

Importance

VRAM is essential for intensive tasks. It serves as the memory for the GPU, allowing it to store and access data quickly. More VRAM can handle larger textures and more complex graphics, which is crucial for high-resolution displays and advanced 3D rendering.

Selection Criteria

- **Graphics Intensity**: More VRAM is needed for graphically intensive tasks such as 3D rendering, gaming, or AI model training that involves large datasets.
- **Parallel Processing Needs**: Tasks that require simultaneous processing of multiple data streams benefit from more VRAM.
- **Future-Proofing**: Opting for more VRAM can make your setup more adaptable to future project requirements.

Storage

Adequate storage, both temporary and persistent, ensures smooth operation and data management.

Importance

Disk size, including both temporary and persistent storage, is critical for data storage, caching, and ensuring that your project has the necessary space for its operations.

Selection Criteria

- **Data Volume**: Estimate the amount of data your project will generate and process.
- **Speed Requirements**: Faster disk speeds can improve overall system performance.
- **Data Retention Needs**: Determine the balance between temporary (volatile) and persistent (non-volatile) storage based on your data retention policies.